IMERG Quality Index

George Huffman

9 November 2017

There have been several requests from users recently for a "simple" quality index to give some guidance on when they should most trust the Integrated Multi-satellitE Retrievals for GPM (IMERG). While the goal is reasonable, there is no agreement about how this quantity should be defined. After some discussion within the team, two distinctly different quality indices were chosen for the half-hourly and monthly data fields for implementation in Version 05. It is a matter of investigation to determine if users find these insightful, or if different quality indices should be developed for future releases.

Quality Index for Half-Hourly Data

At the half-hourly scale, the best metric is some measure of the relative skill that might be expected from the fluctuating mix of different passive microwave- and infrared-based precipitation estimates. The Kalman smoother used in IMERG (and originated in the CPC Morphing [CMORPH] algorithm, Joyce et al. 2011) routinely recomputes estimates of correlation between GMI and each of the other satellite estimates in coarse blocks across the entire IR domain (60°N-S) and then uses these correlation coefficients (squared) to provide weights for use in the combination of forward-propagated passive microwave, backward-propagated passive microwave, and current-time infrared precipitation estimates. However, the formalism never provides an overall correlation for the combined estimate, so one approach is provided here.

The RMS of a combined estimate (σ_t) in terms of the RMS values for the individual estimates (σ_a and σ_b) is given as

$$\sigma_t = \frac{\sigma_a \sigma_b}{\sqrt{\sigma_a^2+\sigma_b^2}} \qquad 1$$

The CMORPH Kalman smoother uses squared correlation coefficient (c^2) in place of $1/\sigma^2$ in the weighting of the input precipitation estimates, so substituting $1/c$ for σ in (1) and simplifying,

$$c_t = \sqrt{c_a^2 + c_b^2}\,, \qquad 2$$

where c_a and c_b are individual correlation coefficients for estimates a and b, and c_t is the estimated correlation coefficient for the combination of estimates a and b.

This formulation has the advantage of producing correlation coefficients higher than the individual input terms, highest when c_a and c_b are equal, and declining to c_a as c_b goes to zero (and vice-versa). However, for both c's close to 1, the resulting c_t can exceed 1 and be as high as 1.414 (square root of 2). One solution to this quandary is to introduce a variance-stabilizing transformation. One simple choice is the Fisher (1915) z statistic

$$z = \operatorname{arctanh}(c)\,, \tag{3}$$

where c is a correlation value. The transformed value z takes on large values as c approaches 1 (or -1), so transforming to z, performing calculations with z, and back-transforming avoids problems around 1. Substituting z for c in (2),

$$c_t = tanh\left(\sqrt{arctanh^2(c_a) + arctanh^2(c_b)}\right) \tag{4}$$

the ordering remains and it gracefully approaches 1. Formally, the Fisher transformation requires that the two variables being correlated follow a bivariate normal distribution. While this is not true for precipitation, we adopt this approach as a first approximation to computing the correlation coefficient of the combined precipitation estimate because it's use as a quality index seems reasonable and useful. In the case of three input correlation coefficients, the equation simply extends to three terms on the right-hand side. The units are non-dimensional correlation coefficients.

There is one additional issue: we lack the zero half-hour correlation of each constellation member to the GMI for computational reasons in the current implementation of IMERG and need an approximate value. Lacking strong justification for alternatives, we chose to set c = 1 when the microwave estimate is present. The next version of this approach should revisit this choice.

Quality Index for Monthly Data

At the monthly scale, a relatively well-founded metric exists for random error, based on Huffman's (1997) analysis of sampling error for a particular data source for a month. The general form of the relationship is

$$\sigma_r^2 = \frac{\bar{r}^2}{N_I}\left(\frac{H}{p} - 1\right), \tag{5}$$

where σ_r is random error, $\bar{r}$ is the time-average of the precipitation rate (originally labeled "rain rate") samples, N_I is the number of independent samples in $\bar{r}$, H is the non-dimensional second moment of the probability distribution of the precipitation rates, and p is the frequency of all nonzero precipitation. Huffman (1997) proceeds to simplify (5) to the approximate expression

$$\sigma_r^2 \cong \frac{H}{I}\frac{(\bar{r}+S)}{N}\left[24 + 49\sqrt{\bar{r}}\right], \tag{6}$$

where $\bar{r}$ and N are available for each grid box in the monthly estimate, I is a multiplicative constant expressing the fraction of N that is "independent", and H/I and S are global constants that are approximated with validation data for each sensor type. This relationship is simple enough that it can be inverted for N. When all the constants are set for the gauge analysis, but the $\bar{r}$ and σ_r^2 used are the final satellite-gauge precipitation estimate and random error variance,

$$N \cong \left(\frac{H}{I}\right)_g \frac{(\bar{r}+S_g)}{\sigma_r^2}\left[24 + 49\sqrt{\bar{r}}\right], \qquad 7$$

and this special *N* is defined as the equivalent number of gauges. Following Huffman (1997), the interpretation is that this is the approximate number of gauges required to produce the estimated random error, given the estimated precipitation. The units are gauges per area, and in the current implementation the area is carried as 2.5°x2.5° of latitude/longitude, even though IMERG is computed on a much finer scale, in order to facilitate interpretation in large-error regions.

Examples

An example of the half-hourly quality index for the IMERG Final Run is shown in Fig. 1. The thin strips of lower quality index are due to gaps between swaths that have not been filled recently. Blockiness is due to the regional variations caused by the coarse resolution and land-ocean separation in the background correlation statistics. Low values at high latitudes are due to two factors. First, microwave estimates are masked out over snowy/icy surfaces, so these regions only have microwave-adjusted IR-based estimates, which have inherently lower correlations. Second, the microwave adjustment to the IR depends on adjustments interpolated from surrounding areas to the areas where microwave estimates have been screened out due to snowy/icy surface. As noted before, grid boxes carrying current-half-hour data from passive microwave input are presently given values of 1, even though the actual correlation should be somewhat lower.

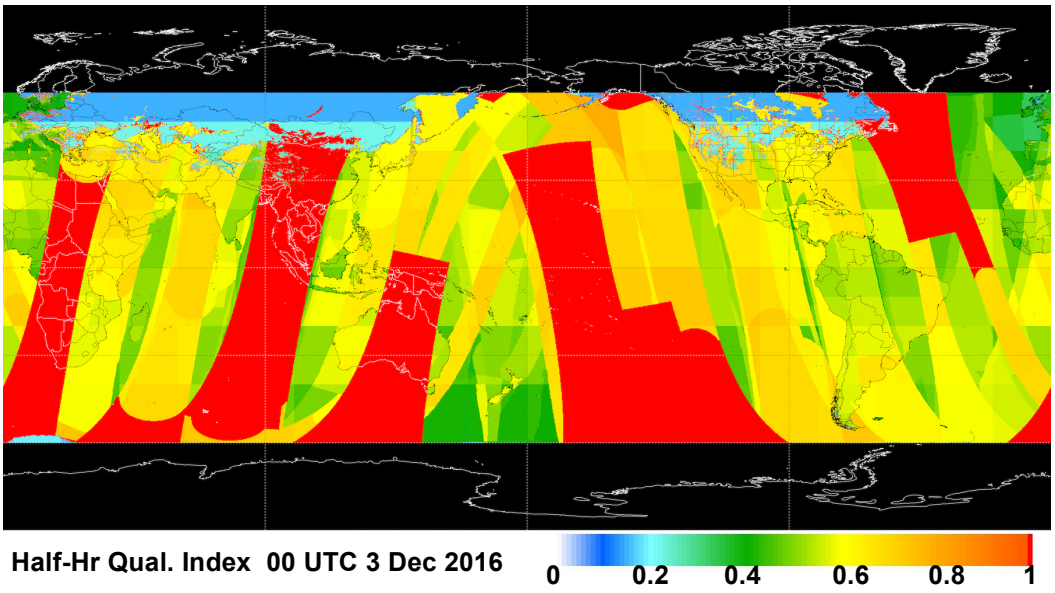

Fig. 1. Quality Index (computed as a composite correlation) for the half-hourly IMERG Final Run for the period 0000-0030 UTC on 3 December 2016. Blacked-out areas lack data. [Courtesy D. Bolvin (SSAI; GSFC)]

An example of the monthly quality index for the IMERG Final Run is shown in Fig. 2. [Recall that only the Final Run has monthly data as a native product.] Over oceans, the equivalent gauges metric largely tames the variation of random error with precipitation rate. Over land, the index largely reflects the distribution of precipitation gauges, except it has a floor of satellite estimates (similar to the values over ocean) where gauges are extremely sparse. The values outside the morphing region (60·N-S) reflect relatively sparse gauges (over snowy/icy land) and passive microwave sampling over ice-free ocean.

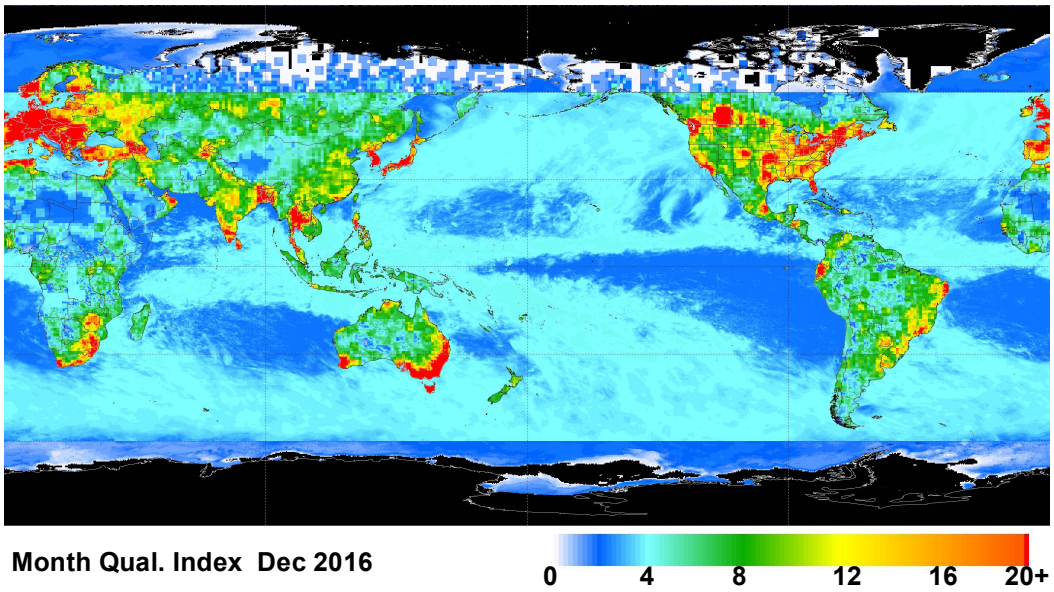

Fig. 2. Quality Index (computed as equivalent gauges per 2.5°x2.5° lat./lon. box) for the Monthly IMERG Final Run for December 2016. Blacked-out areas lack data. [Courtesy D. Bolvin (SSAI; GSFC)]

References

Fisher, R.A., 1915: Frequency Distribution of the Values of the Correlation Coefficient in Samples of an Indefinitely Large Population. *Biometrika*, Biometrika Trust, **10**(4), 507–521. *doi:10.2307/2331838*

Huffman, G.J., R.F. Adler, D.T. Bolvin, G. Gu, E.J. Nelkin, K.P. Bowman, Y. Hong, E.F. Stocker, D.B. Wolff, 2007: The TRMM Multi-satellite Precipitation Analysis: Quasi-Global, Multi-Year, Combined-Sensor Precipitation Estimates at Fine Scale. *J. Hydrometeor.*, **8**, 38-55. *doi:10.1175/JHM560.1*

Joyce, R.J., P. Xie, J.E. Janowiak, 2011: Kalman Filter Based CMORPH. *J. Hydrometeor.*, **12**, 1547-1563. *doi:10.1175/JHM-D-11-022.1*